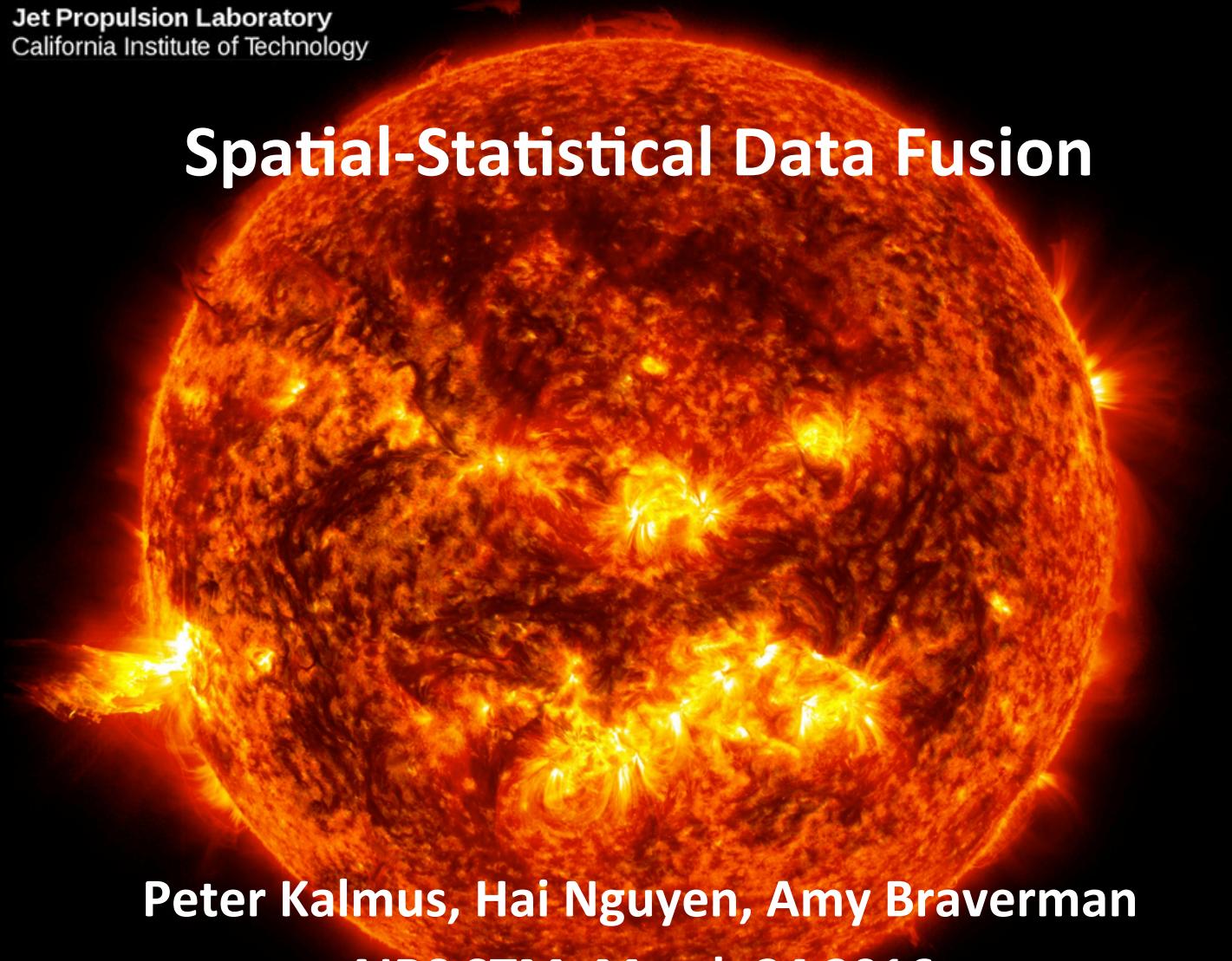




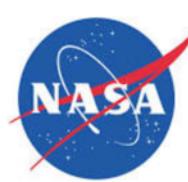
© 2016 California Institute of Technology.
Government sponsorship acknowledged.

Spatial-Statistical Data Fusion

A large, detailed image of the Sun's surface, showing its granular texture and several bright, explosive solar flares erupting from various points across its visible disk.

Peter Kalmus, Hai Nguyen, Amy Braverman

AIRS STM. March 24 2016



National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Motivation

Suppose there are i data sets estimating the same field.

Different spatio-temporal support; different strengths & weaknesses.

Example: global satellite temperature soundings.

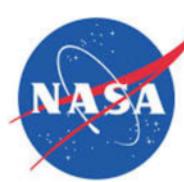
Why not fuse the i data sets into one optimal estimate?

Optimal = unbiased, minimal variance

And also produce robust uncertainty estimates?

We present an early proof-of-concept for such a data fusion.





Spatial-Statistical Data Fusion

Start with 2+ noisy, incomplete data sets representing some physical field (e.g. T, q).

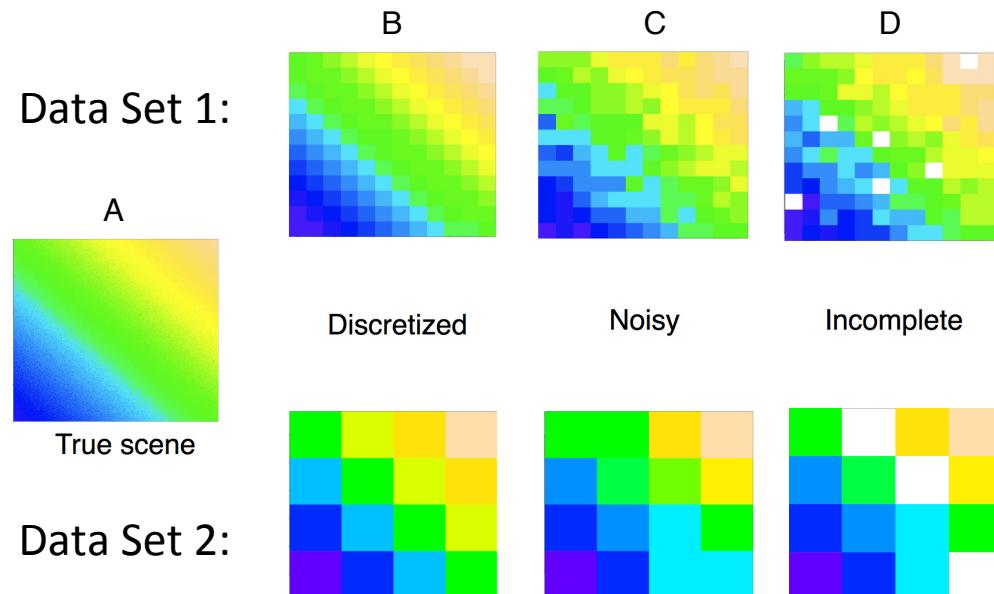
Each data set must come with robust bias and variance estimates. Spatial footprints can be anything.

We model the underlying (true) physical field as a spatial Gaussian process.

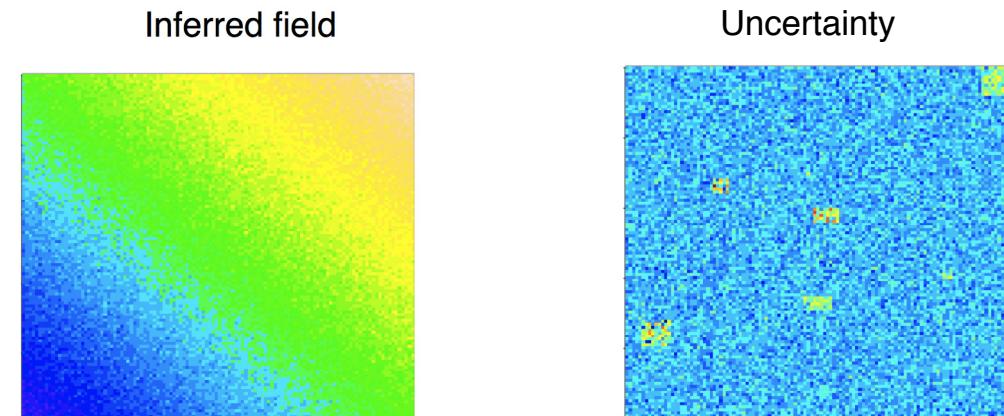
We model the observations as functions of these random variables (e.g., spatial aggregates plus measurement error).

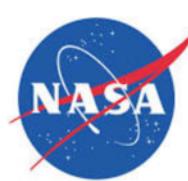
From the observations, we estimate the spatial covariance of the underlying field and infer the true field, accounting for spatial dependence and error of each input data set.

for details, see Nguyen, Cressie, and Braverman 2012



Fusion of column D yields:





Evolution of fusion methods

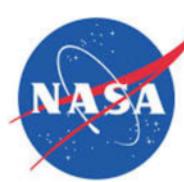
Two major challenges:

- different spatial footprints of input data sets
 - Bayesian melding (e.g. Fuentes & Raftery 2005)
- massive data sets of size N
 - Bayesian melding has computational complexity $O(N^3)$
 - Fixed-Rank Kriging (FRK), single dataset spatial interpolation $O(N)$
(Cressie & Johannesson 2008)

SSDF generalizes FRK to multiple data sets

Different areal footprints, measurement errors

Performance comparable to Bayesian melding, but much faster.



Proof-of-Concept

Fusion of 3 data sets: AIRS+ECMWF+MERRA2

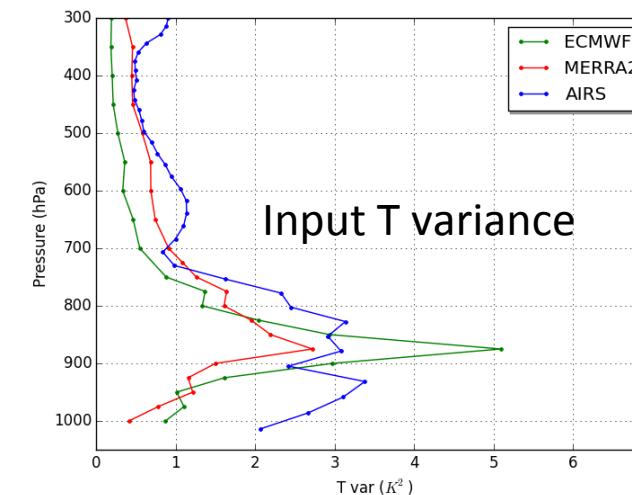
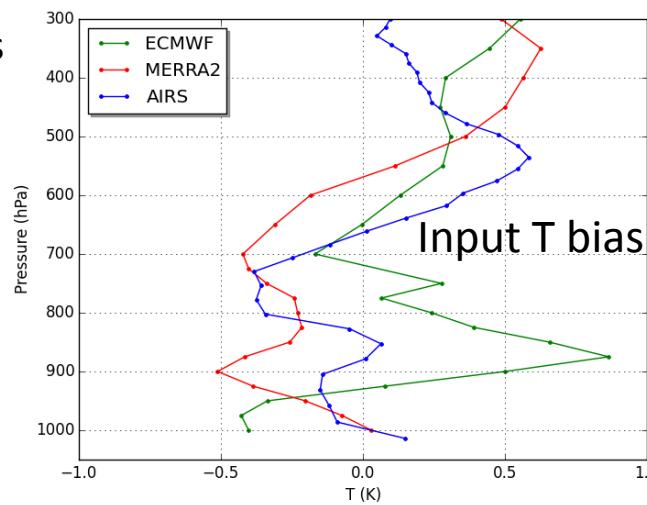
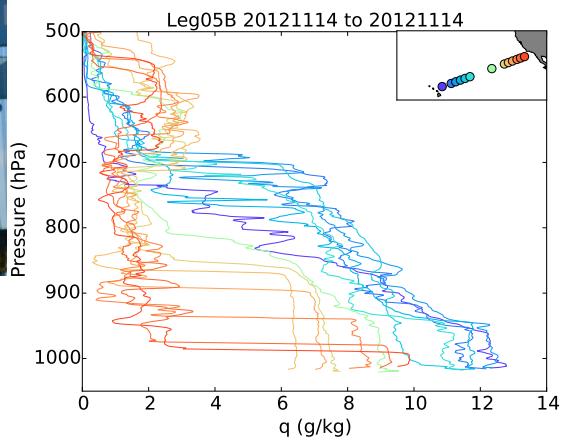
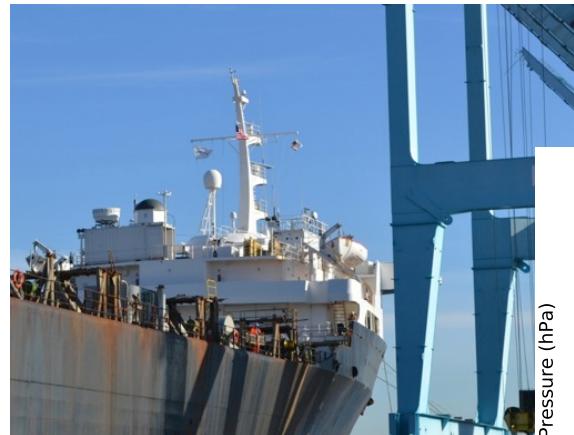
- levels fused separately
- T, q fused separately (for now)
- **July 2013, 10N–40N, 130W–160W**

MAGIC radiosondes:

- estimate bias, variance of input data sets (over entire MAGIC campaign)
- evaluate results (July only)
- matchups: 200km, ± 6 hours

Fused data output:

- 6-hourly, 0.5 degree grid
- q, T on 14 pressure levels

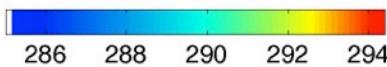
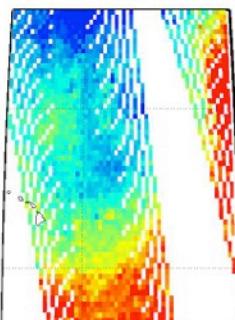




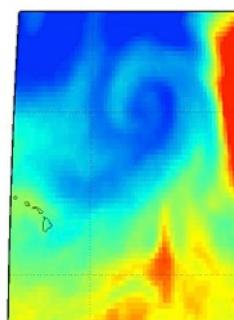
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Preliminary results: T, 925 hPa

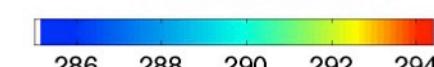
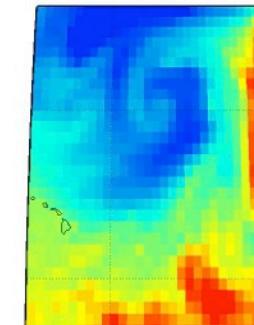
AIRS



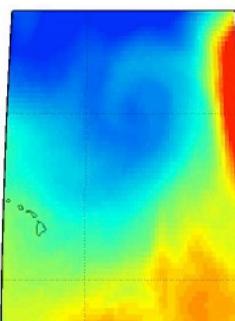
ECMWF ERA-Interim



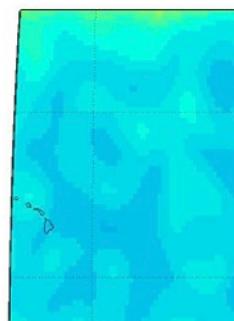
MERRA2



Fused

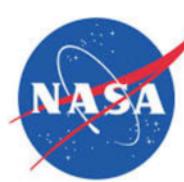


Variance of fused



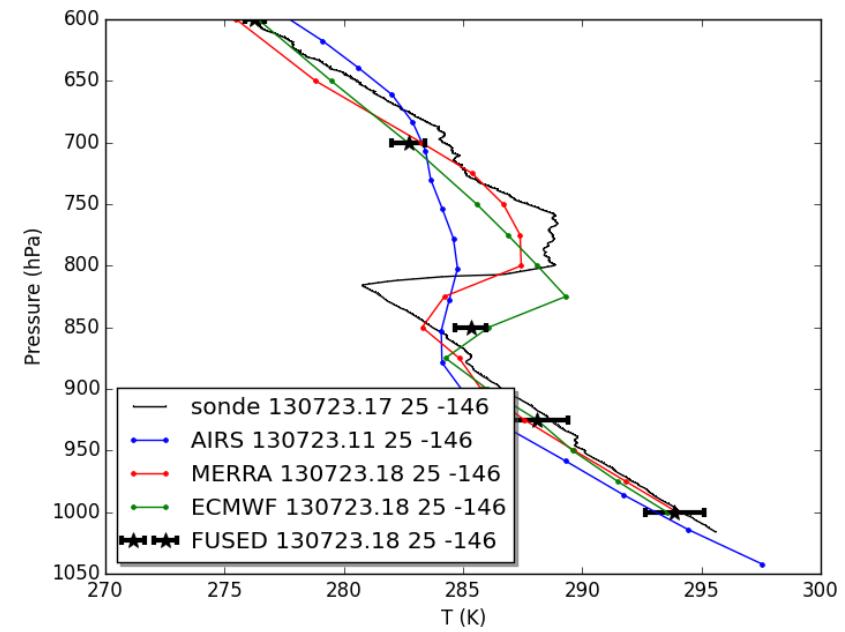
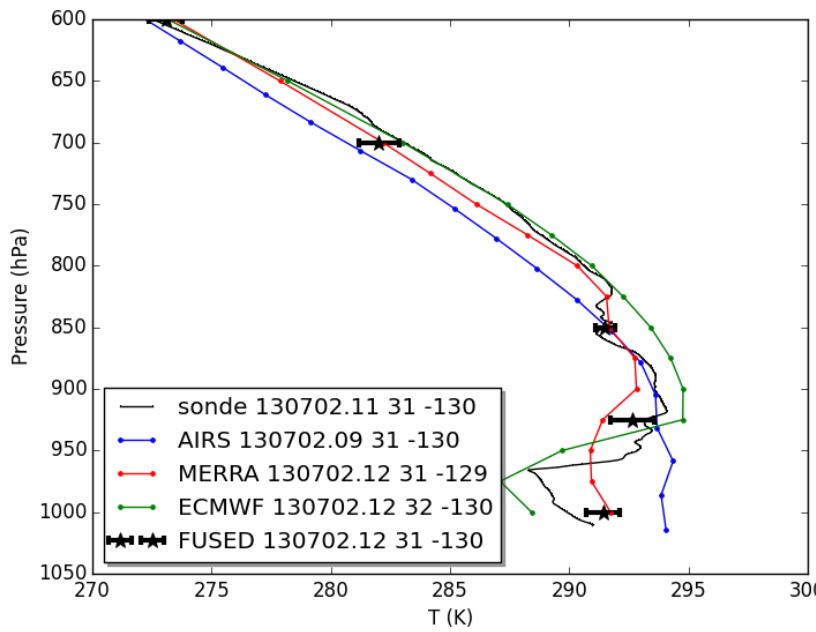
Example: 925 hPa temperature (K)
July 2 2013, 1800 UTC

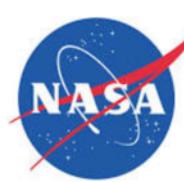
Fusing 3 sparse data sets would further highlight the usefulness of fusion.



National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

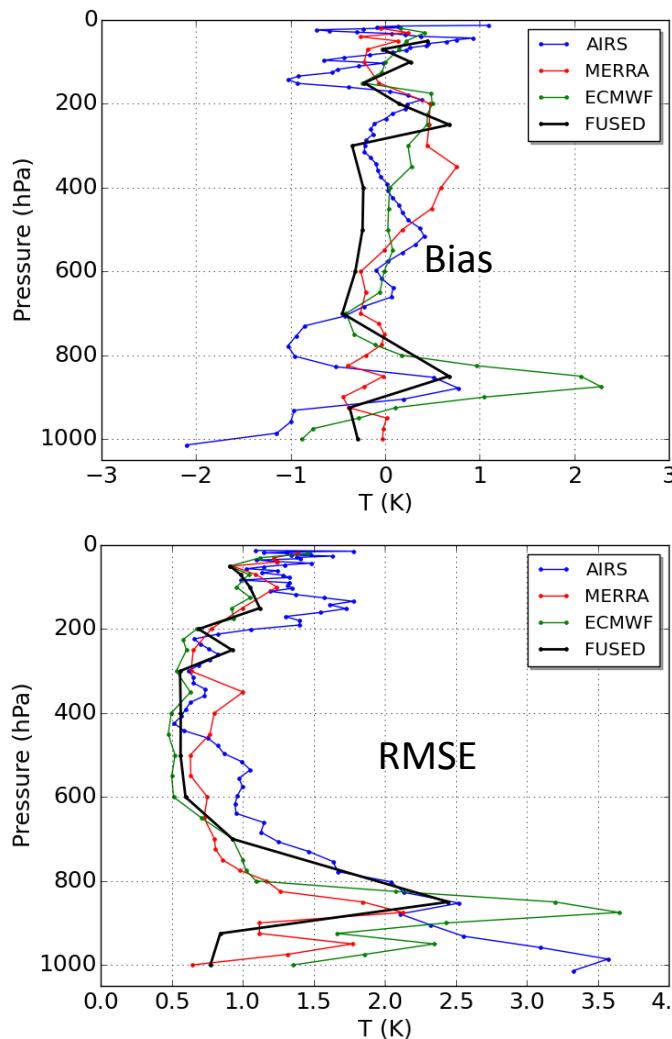
Example T profiles



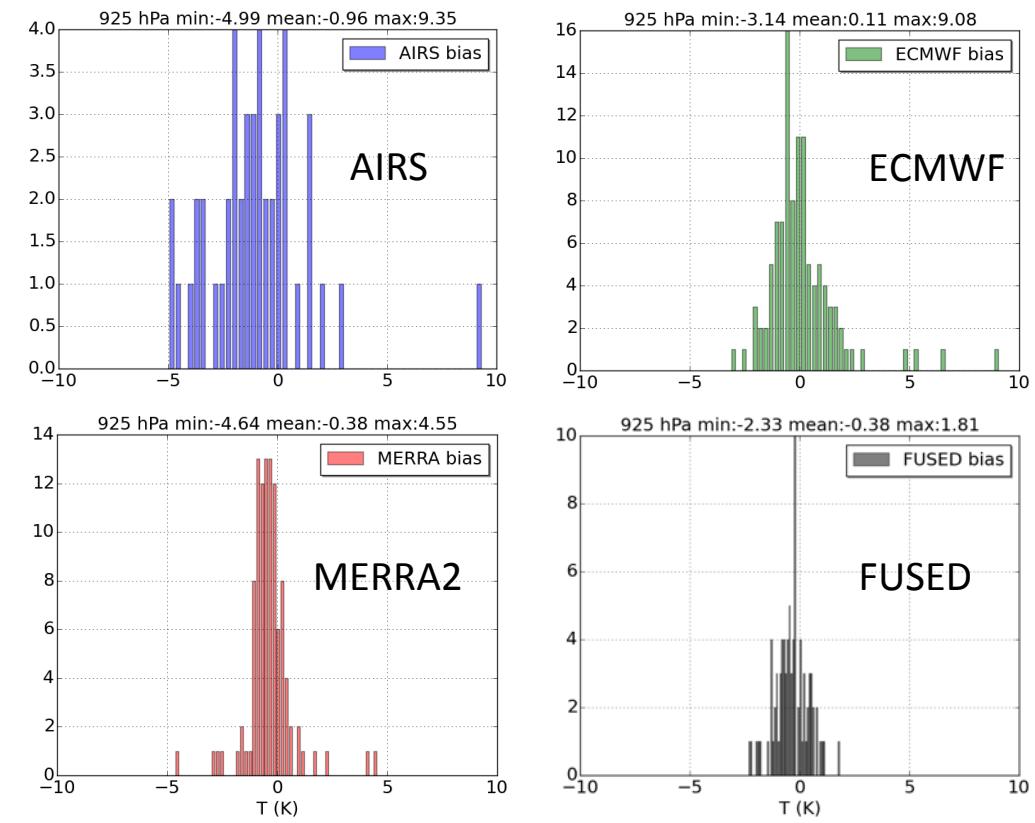


National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

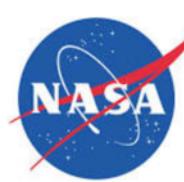
Preliminary results: T



Above: Profiles of T mean bias, RMSE



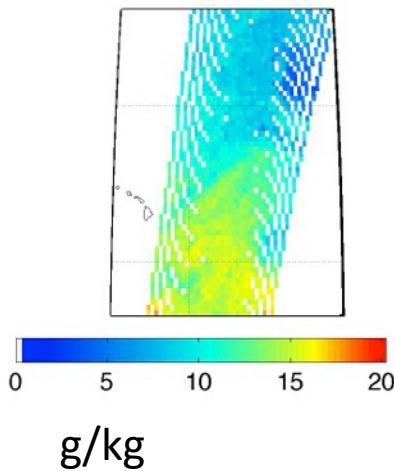
Above: 925 hPa T bias histograms



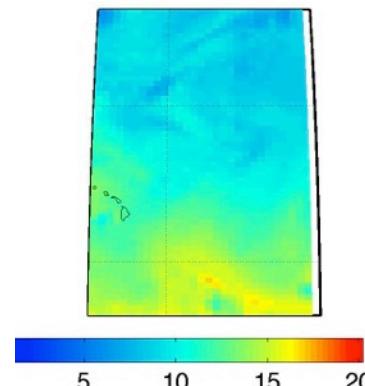
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Prelim. results: q, 925 hPa

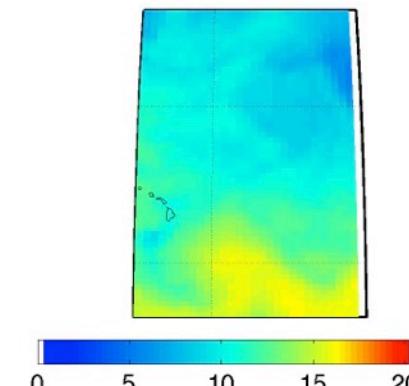
AIRS



ECMWF ERA-Interim

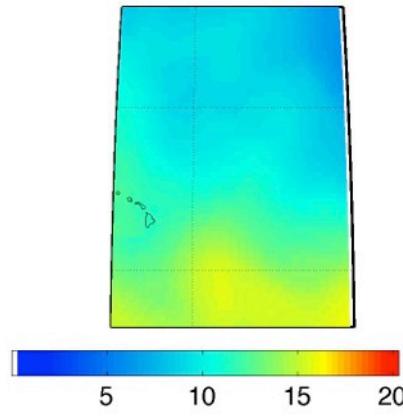


MERRA2

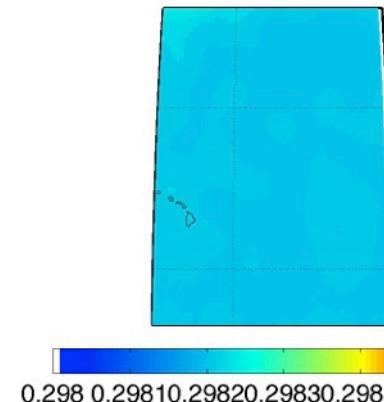


g/kg

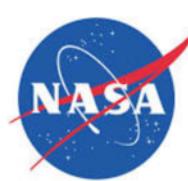
Fused



Variance of fused

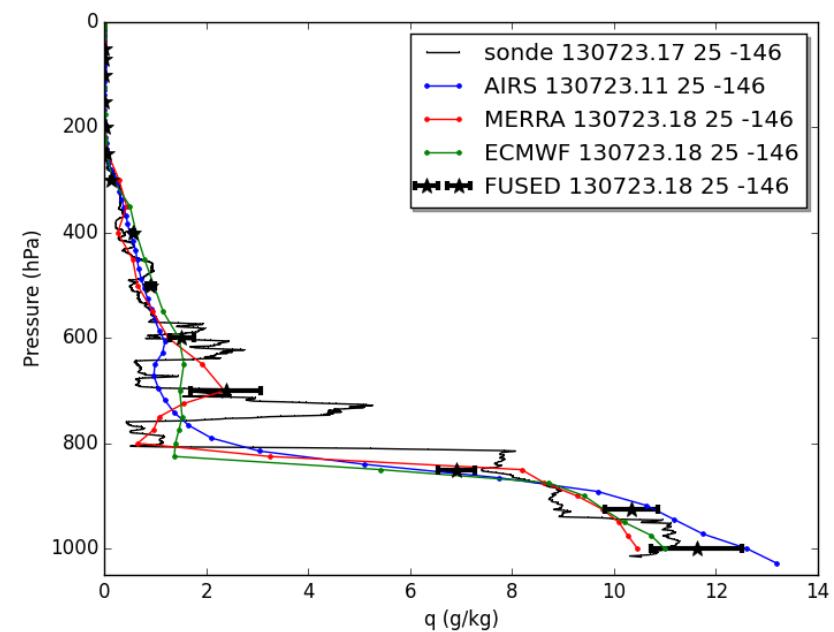
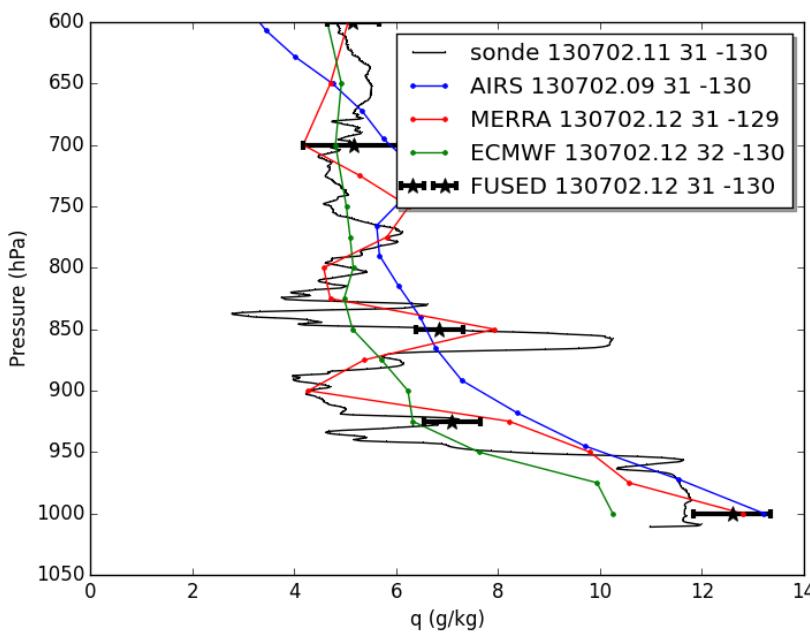


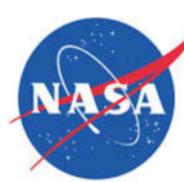
Example: 925 hPa q (g/kg)
July 19 2013, 0600 UTC



National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

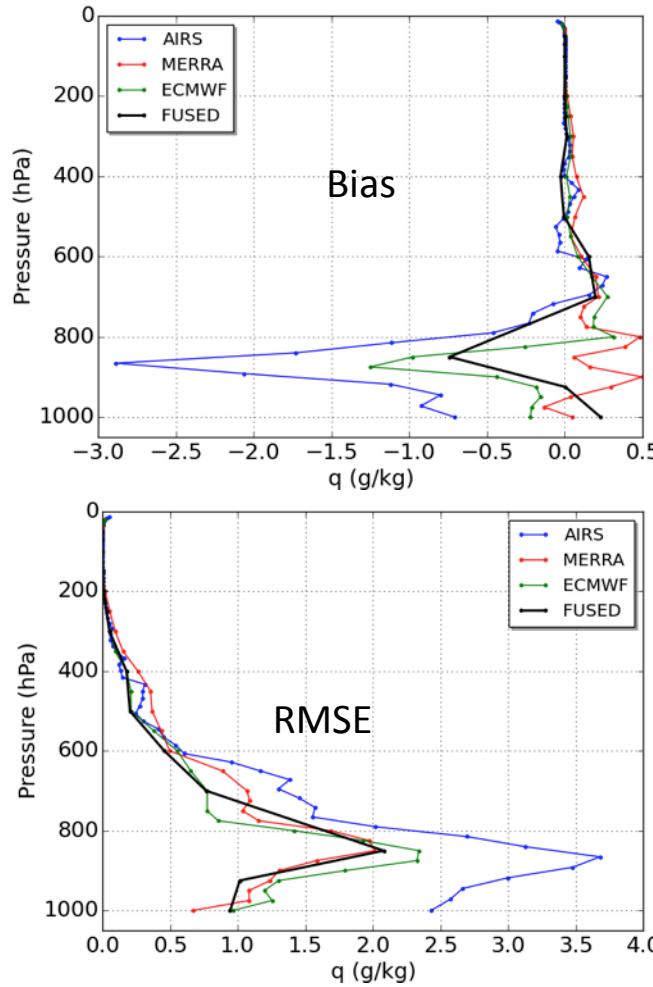
Example q profiles



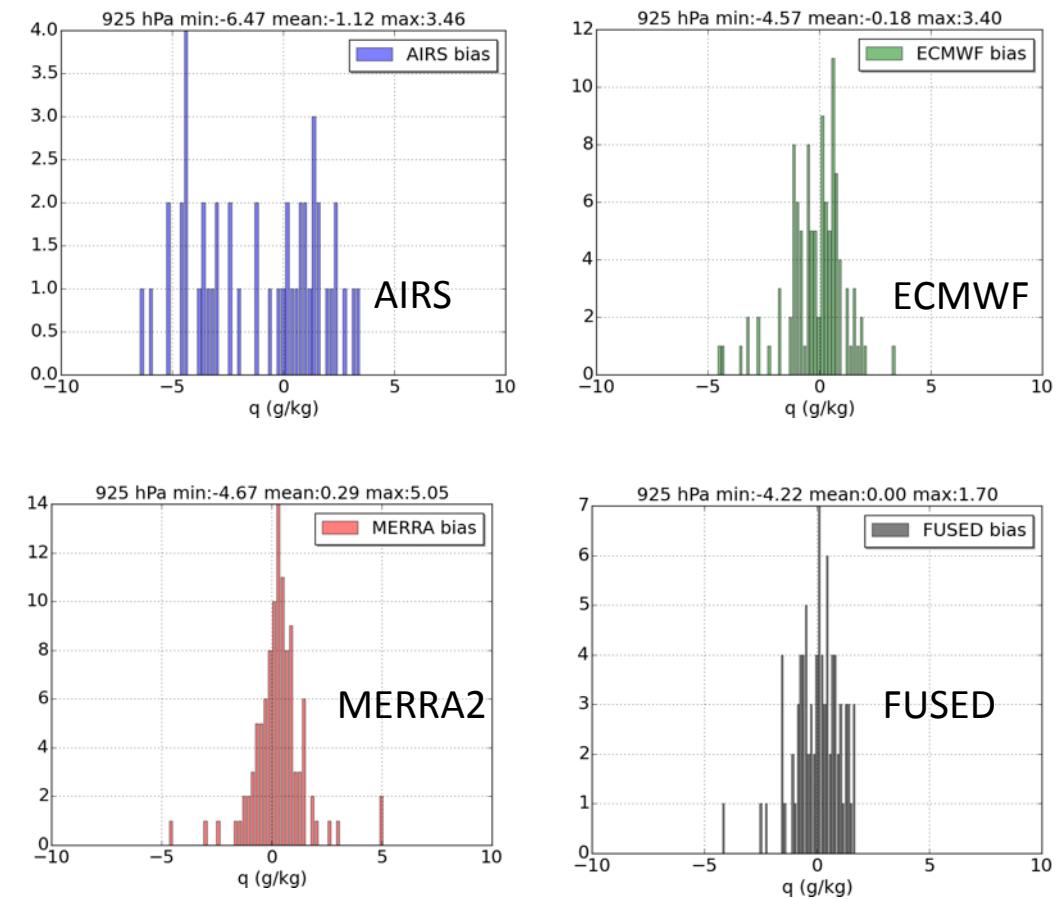


National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

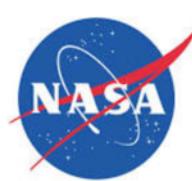
Preliminary results: q



Above: Profiles of q mean bias, RMSE



Above: 925 hPa q bias histograms

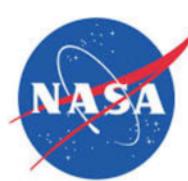


Error estimates & synthetic data

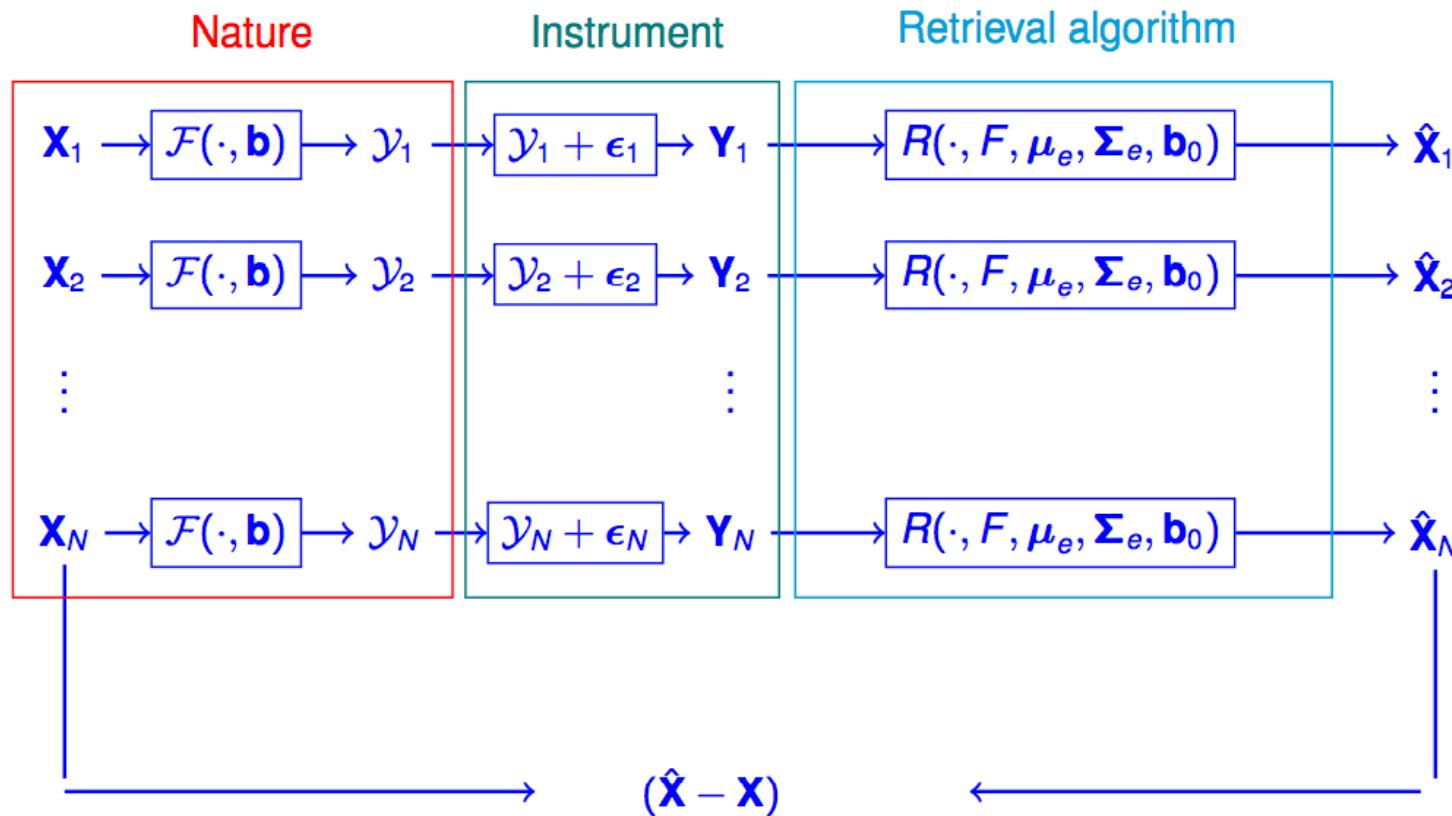
We used MAGIC to estimate error for the proof-of-concept.
But it won't suffice for a global data fusion.

How to get robust global bias and variance estimates for global input data sets?

One possibility: characterize uncertainty in retrievals with simulations using synthetic “truth” data.



Synthetic error estimation



Would need to do this per regime.

Not perfect, therefore not necessarily the enemy of the good.

Conclusions

Fused data sets could be valuable to the community.
SSDF is a promising method for fusing data sets.

Next steps:

Simultaneous T, q fusion (AIRS+ECMWF+MERRA2)

- quantify improvement

AIRS+CrIS+IASI T, q fusion in “MAGIC sandbox” (NE Pacific 30°x30°)

Optimize SSDF performance

- error calculation currently takes 50—90% of CPU time
- 1 yr of global AIRS+ECMWF+MERRA would take 60 hours on 5000 CPUs

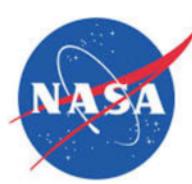
Global estimates of AIRS T, q bias and variance using synthetic data

Global proof-of-concept for input data sets (AIRS+CrIS+IASI T,q)



National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Backup slides



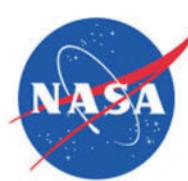
Method 1

We assume the data are generated according to the following model:

$$\begin{aligned}\mathbf{Z} &= (Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_N))', \\ Z(\mathbf{s}) &= Y(\mathbf{s}) + \epsilon(\mathbf{s})\end{aligned}$$

where

- ▶ \mathbf{s}_i is the i th footprint ,
- ▶ \mathbf{Z} is the vector of response variable,
- ▶ $Y(\cdot)$ is the true constant-mean process
- ▶ $\epsilon(\cdot)$ is the error process.



Method 2

Under this formulation, the (linear unbiased) optimal interpolation can be written as

$$\hat{Y}(\mathbf{s}) = \mathbf{a}' \mathbf{Z}$$

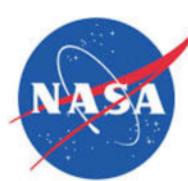
where \mathbf{a} is a N -dimensional vector of *kriging coefficients* at location \mathbf{s} .

We wish to find the vector \mathbf{a} that minimizes,

$$\begin{aligned} E(Y(\mathbf{s}) - \hat{Y}(\mathbf{s}))^2 &= \text{var}(Y(\mathbf{s}) - \mathbf{a}' \mathbf{Z}), \\ &= \text{var}(Y(\mathbf{s})) - 2\mathbf{a}' \text{cov}(\mathbf{Z}, Y(\mathbf{s})) + \mathbf{a}' \text{var}(\mathbf{Z}) \mathbf{a}, \end{aligned} \quad (1)$$

with respect to \mathbf{a} , subject to the unbiasedness constraint,

$$\mathbf{1} = \mathbf{a}' \mathbf{1},$$



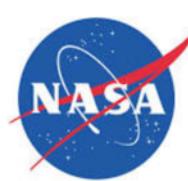
Method 3

We can solve (1) for the optimal \mathbf{a} using the method of Lagrange multiplier. The equation for the optimal kriging coefficients is,

$$\begin{pmatrix} \Sigma & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{c}(\mathbf{s}) \\ 1 \end{pmatrix} \quad (2)$$

where

- ▶ $\Sigma \equiv \text{var}(\mathbf{Z})$
- ▶ $\mathbf{c}(\mathbf{s}) \equiv \text{cov}(\mathbf{Z}, Y(\mathbf{s}))$
- ▶ λ is the Lagrange multiplier



Method 4: Fixed Rank Kriging

Note that to solve Equation (2), we need to invert the $(N + 1) \times (N + 1)$ matrix

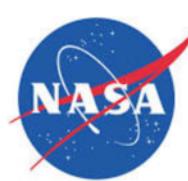
$$\begin{pmatrix} \Sigma & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}.$$

The above could easily be done using inversion by block, but we do need to invert the smaller $N \times N$ matrix Σ .

In general, inversion of an $N \times N$ matrix has computational complexity $O(N^3)$.

Fixed Rank Kriging was designed to resolve the scalability issue.

It works by assuming a certain structure on $Y(\cdot)$, which leads to quick inversion of the covariance matrix Σ .



Method 5: FRK

We assume that the spatial process $Y(\cdot)$ has the following model,

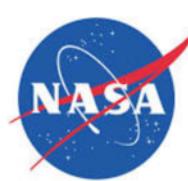
$$Y(s) = S(s)' \eta,$$

which leads to the following covariance model,

$$\Sigma \equiv \text{var}(Z) = S' K S + D,$$

where

- ▶ $S(s)$ is an r -dimensional basis expansion of s , and $r \ll N$,
- ▶ $S \equiv (S(s_1), \dots, S(s_N))'$,
- ▶ $K = \text{var}(\eta)$: fixed dimension $r \times r$,
- ▶ D is the variance-covariance matrix of the measurement errors.



Method 6: FRK

Since Σ has the convenient form

$$\Sigma \equiv \text{var}(\mathbf{Z}) = \mathbf{S}' \mathbf{K} \mathbf{S} + \mathbf{D},$$

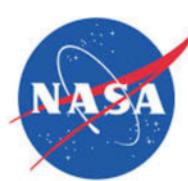
We can quickly invert Σ using the Sherman-Morrison-Woodbury formula

$$\Sigma^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1} \mathbf{S}' (\mathbf{K}^{-1} + \mathbf{S} \mathbf{D}^{-1} \mathbf{S}')^{-1} \mathbf{S} \mathbf{D}^{-1}. \quad (3)$$

From Equation (3), we need to invert three terms

- ▶ \mathbf{D} : a diagonal $N \times N$ matrix,
- ▶ \mathbf{K} : an $r \times r$ matrix,
- ▶ $(\mathbf{K}^{-1} + \mathbf{S} \mathbf{D}^{-1} \mathbf{S}')$: also an $r \times r$ matrix.

The computational complexity is $O(Nr^2)$, which is linear with respect to the data size N .



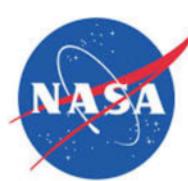
Method 7: Multiple data sets

We assume the data from instrument i are generated according to the following model:

$$\begin{aligned}\mathbf{Z}_i &= (Z_i(\mathbf{s}_{i1}), Z_i(\mathbf{s}_{i2}), \dots, Z_i(\mathbf{s}_{iN_i}))', \\ Z_i(\mathbf{s}_{ij}) &= Y(\mathbf{s}_{ij}) + \epsilon_i(\mathbf{s}_{ij}); \\ Y(\mathbf{s}_{ij}) &= \mathbf{S}(\mathbf{s}_{ij})\boldsymbol{\eta};\end{aligned}$$

where

- ▶ \mathbf{Z}_i is the vector of response variable from dataset i ,
- ▶ $Y(\cdot)$ is the true process,
- ▶ $\epsilon_i(\mathbf{s}_{ij})$ is the error process.



Method 8: Multiple data sets

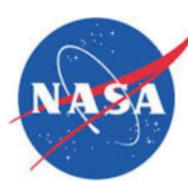
We can rewrite the data models in vector forms as

$$\mathbf{Z}_1 = \mathbf{S}'_1 \boldsymbol{\eta} + \boldsymbol{\epsilon}_1$$

$$\mathbf{Z}_2 = \mathbf{S}'_2 \boldsymbol{\eta} + \boldsymbol{\epsilon}_2$$

where

- ▶ \mathbf{Z}_i is the vector of response variable from dataset i ,
- ▶ $\boldsymbol{\eta}$ is the **hidden process common to both dataset**,
- ▶ $\boldsymbol{\epsilon}_i(B_{ij})$ is the error process.



Method 9: Multiple data sets

Given the data vectors \mathbf{Z}_1 and \mathbf{Z}_2 , we can simplify the problem by stacking them to form the following model,

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{S}'_1 \\ \mathbf{S}'_2 \end{pmatrix} \boldsymbol{\eta} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix},$$

or equivalently,

$$\mathbf{Z}_F = \mathbf{S}' \boldsymbol{\eta} + \boldsymbol{\epsilon}_F.$$

Under this formulation, the (linear unbiased) optimal data fusion can be written as

$$\begin{aligned}\hat{Y}(\mathbf{s}) &= \mathbf{a}'_1 \mathbf{Z}_1 + \mathbf{a}'_2 \mathbf{Z}_2, \\ &= \mathbf{a}'_F \mathbf{Z}_F,\end{aligned}$$

where $\mathbf{a}'_F \equiv (\mathbf{a}'_{1s}, \mathbf{a}'_{2s})$.

Note that we now have a single meta-dataset, and the solution for \mathbf{a}_F is precisely the same as the 1-dataset case.